Applied Analytics Practicum - Enverus

**Team 5**

hdupouy3@gatech.edu  
oalmansour3@gatech.edu  
jgarcia340@gatech.edu

***Abstract -*** This project presents a study to evaluate the impact of various locations and modeling techniques on predicting solar farm energy generation's performance. This research project was undertaken in collaboration with Enverus, an energy technology company renowned for its innovative software, data, and services in the energy market. This study compares and contrasts different modeling options by incorporating location-specific source data and alternative modeling techniques. This paper discusses the investigation's problem statement, objectives, and desired goals towards solar energy generation.

Table of Contents

[1. Introduction 4](#_Toc139818891)

[1.1. Enverus 4](#_Toc139818892)

[1.2. Objective 4](#_Toc139818893)

[1.2.1. Project objectives 4](#_Toc139818894)

[1.2.1. Value and Magnitude 4](#_Toc139818895)

[1.3. Problem statement 5](#_Toc139818896)

[1.4. Hypotheses 6](#_Toc139818897)

[2. Literature review 7](#_Toc139818898)

[3. Methodology 8](#_Toc139818899)

[3.1. Data Collection 8](#_Toc139818900)

[3.2. Preprocessing 9](#_Toc139818901)

[3.3. Data Modeling 10](#_Toc139818902)

[3.4. Experimentation Design 11](#_Toc139818903)

[4. Explanatory Data Analysis 13](#_Toc139818904)

[4.1. General trends 13](#_Toc139818905)

[4.2. Feature selection 17](#_Toc139818906)

[5. Results and Visualizations 21](#_Toc139818907)

[5.1. Linear Regression Modeling 21](#_Toc139818908)

[5.2. Random Forest 24](#_Toc139818909)

[5.3. Time Series Modeling 26](#_Toc139818910)

[5.4. XGBoost 29](#_Toc139818911)

[6. Conclusion and discussion 34](#_Toc139818912)

[7. References 35](#_Toc139818913)

[8. Appendix A 37](#_Toc139818914)

Table of Figures

[Figure 1. ETL Data Pipeline 8](#_Toc139818854)

[Figure 2. Actual value by hour over the entire period 13](#_Toc139818855)

[Figure 3. Actual value by hour through the day 14](#_Toc139818856)

[Figure 4. Comparing a week of Actual value in summer and winter 15](#_Toc139818857)

[Figure 5. Diagram of the analemma 15](#_Toc139818858)

[Figure 6. Representation of the 15](#_Toc139818859)

[Figure 7. Diagram of the analemma 16](#_Toc139818860)

[Figure 8. Diagram of the analemma 16](#_Toc139818861)

[Figure 9. First Correlation Matrix 17](#_Toc139818862)

[Figure 10. Final Correlation Matrix 18](#_Toc139818863)

[Figure 11. Density of the numerical variables 19](#_Toc139818864)

[Figure 12. Distribution of the categorical variables 20](#_Toc139818865)

[Figure 13. Prediction vs Actual; & error distribution for simple linear regression 21](#_Toc139818866)

[Figure 14. LR - Actual and model power generation in a winter week in 2021 & 2022 22](#_Toc139818867)

[Figure 15. LR - Actual and model power generation in a summer week in 2021 & 2022 22](#_Toc139818868)

[Figure 16. LR scaled - Prediction vs Actual; & error distribution for MLR model 23](#_Toc139818869)

[Figure 17. LR scaled - Actual and model power generation in a winter week in 2021 & 2022 23](#_Toc139818870)

[Figure 18. LR scaled - Actual and model power generation in a summer week in 2021 & 2022 24](#_Toc139818871)

[Figure 19. RF - RMSE & R2 vs Number of Trees 25](#_Toc139818872)

[Figure 20. RF - Prediction vs Actual; & error distribution for random forest model 25](#_Toc139818873)

[Figure 21. RF - Actual and model power generation in a summer week in 2021 & 2022 26](#_Toc139818874)

[Figure 22. RF - Actual and model power generation in a summer week in 2021 & 2022 26](#_Toc139818875)

[Figure 23. ARIMA - Actual and model power and Autoregressive partial autocorrelation 27](#_Toc139818876)

[Figure 24. ARIMA - Actual and model power generation in a summer week in 2021 & 2022 27](#_Toc139818877)

[Figure 25. Autoregressive - Autoregressive model performance in 2 different weeks 28](#_Toc139818878)

[Figure 26. XGB - Feature importance on non-optimized model 29](#_Toc139818879)

[Figure 27. XGB Optuna - Optimization History Plot 29](#_Toc139818880)

[Figure 28. XGB Optuna - Parallel Coordinate Plot 30](#_Toc139818881)

[Figure 29. XGB Optuna - Hyperparameter Importance 30](#_Toc139818882)

[Figure 30. XGB Optimized - Feature importance on optimized model 31](#_Toc139818883)

[Figure 31. XGB Optimized - Model Predictions vs Actual 31](#_Toc139818884)

[Figure 32. XGB Optimized - Model Predictions vs Actual focus on predictions period 32](#_Toc139818885)

[Figure 33. XGB Optimized - Model Predictions vs Actual 32](#_Toc139818886)

# Introduction

## 1.1. Enverus

Enverus, an energy technology company, is the driving force behind this research endeavor. Since its establishment in 1999, Enverus has emerged as a prominent provider of energy market data, analytics, and technology solutions (Enverus, 2023). With a commitment to optimizing operations and fostering a deeper understanding of energy markets, Enverus offers innovative software, data, and services to facilitate informed decision-making within the energy sector. Indeed, Enverus gives energy firms the platforms, tools, and applications they need to be adaptable and thrive in a challenging and changing market environment. Additionally, Enverus provides invaluable services such as expert guidance, data analysis, and market intelligence, further solidifying its position as a leader in the field.

## 1.2. Objective

### 1.2.1. Project objectives

This project's primary objective is to study the effect of incorporating location-specific source data and various modeling techniques to predict energy generation performance from solar farms. We will compare and contrast several models in an effort to improve energy generation in a specific location. This will be achieved through extensive research and experimentation. Enverus will provide us with an anonymized dataset. This dataset contains the target variable and timestamps extracted from California ISO toward the actual solar megawatt generation (California ISO, 2023). This research will contribute to the existing literature on the prediction of the energy generation performance of solar farms, and it will also assist energy companies like Enverus in making informed decisions in order to thrive in the ever-changing energy market.

### 1.2.1. Value and Magnitude

This project holds substantial value for the company by generating valuable insights and knowledge concerning the prediction of energy generation. By integrating location-specific source data and advanced modeling techniques, the project aims to enhance solar farm performance models' accuracy and predictive performance. By improving the precision and reliability of these models, energy companies can make more informed decisions regarding their operations, resource allocation, and identification of regions with high solar energy generation potential. This optimization of processes and resource utilization can lead to cost savings and overall performance improvements, enabling companies to stay competitive in a rapidly changing market.

The magnitude of this project is noteworthy due to its comprehensive approach and incorporation of multiple elements. This project requires external data and time-series forecasting models to produce predictions. Based on the combination of the data provided and retrieved, it needs to leverage machine learning models and visualization libraries. Indeed, the scale of the project involves analyzing a substantial amount of data, with up to 50,000 rows and 20 input variables per data source. Furthermore, advanced modeling techniques, such as Random Forest and eXtreme Gradient Boosting (XGBoost), capture complex relationships and patterns within the data. The project also includes a thorough review of relevant literature on solar generation forecasting to incorporate best practices and industry knowledge. Given the comprehensive nature of the project and its potential impact on decision-making processes within energy companies, the magnitude of this endeavor is significant.

## 1.3. Problem statement

Existing methods for forecasting solar farm performance have limitations due to a lack of consideration for location-specific data and reliance on traditional modeling tools. This leads to inaccurate and unreliable predictions. To address this issue and improve the accuracy and foresight of solar farm performance models, evaluating and comparing the outcomes obtained by utilizing location-specific data sources and advanced modeling methodologies is necessary. By doing so, energy providers can enhance operational efficiency, optimize resource allocation, and identify areas with the highest potential for solar power production.

Research questions:

• When comparing generic data sources and traditional modeling methods to solar energy-generating performance prediction, how does integrating location-specific data sources and sophisticated modeling techniques affect the accuracy and predictive performance?

• Moreover, how might these enhancements help energy firms maximize operations, allocate resources, and pinpoint areas with the most significant potential for solar power generation?

## 1.4. Hypotheses

Our first hypothesis mentioned that location and modeling methods affect solar farm performance. Using location-specific source data, which accounts for physical closeness to the place under investigation, the models should reflect various locales' distinctive traits and variances. Solar irradiance, weather patterns, topographical characteristics, and environmental factors may be included. This study's design matrix includes data from nearby macro, meso, and microregions. This study aims to solve research questions and understand solar generation performance aspects.

Indeed, this hypothesis also predicts that sophisticated algorithms like Random Forest, XGBoost, and others can capture complicated linkages and nonlinearities in data. These methods may reveal patterns and relationships that linear regression models miss. By merging advanced modeling approaches with location-specific data sources, solar farm performance projections will be more precise and dependable. This helps energy businesses optimize operations, allocate resources, and find solar energy-generating zones.

The study will investigate numerous data sources and model classes to test our notion. If considerable gains are found, the models' accuracy, robustness, and generalizability will demonstrate the suggested approach's usefulness.

# Literature review

In our specific study, the geographical location under investigation is within the California region. However, it is important to acknowledge that the exact geographical location within California plays an important role in the performance of solar farms, as pointed out by Miao, Ning, Gu, Yan, and Ma (2018). These authors highlight the variability of solar radiation across different locations and propose a comprehensive framework that considers factors such as latitude, longitude, and weather patterns to assess their impact on energy generation. The results of their research highlight significant differences in performance depending on the specific geographical location, further emphasizing the need for accurate modeling techniques (Miao, Ning, Gu, Yan, & Ma, 2018). Therefore, for our project, we will strive to identify a precise location within California that allows for robust data analysis and the use of valued data science techniques for predicting time series data.

Solar power generation is influenced by several environmental factors, as highlighted by Singh and Singh (2021). These factors include solar irradiance, temperature, humidity, dust, shading, and wind speed. In addition, the technical design characteristics of photovoltaic cells, such as the materials used in their manufacture, also affect the power generation capabilities of the cells (Chikate & Sadawarte, 2015). Therefore, we don't have access to photovoltaic cell specifications in our data collection process. However, we aim to collect additional data points in order to obtain a comprehensive overview of all significant factors affecting the performance of solar energy generation.

Similar studies in the field of solar energy generation have consistently emphasized the need for robust modeling techniques. Hobbs et al. (2022), in their research of probabilistic solar prediction using the probabilistic Watt-Sun model, underlined the need for a reliable and accurate model despite the inherent uncertainties associated with prediction predictors. Similarly, Aksoy and Genc (2023) proposed an ensemble model that combines several basic models, such as Random Forests and Gradient Boosting Machines, to improve the accuracy of energy production forecasts. Their research demonstrates the effectiveness of ensemble learning in capturing complex relationships between input variables and energy production (Aksoy & Genc, 2023).

# Methodology

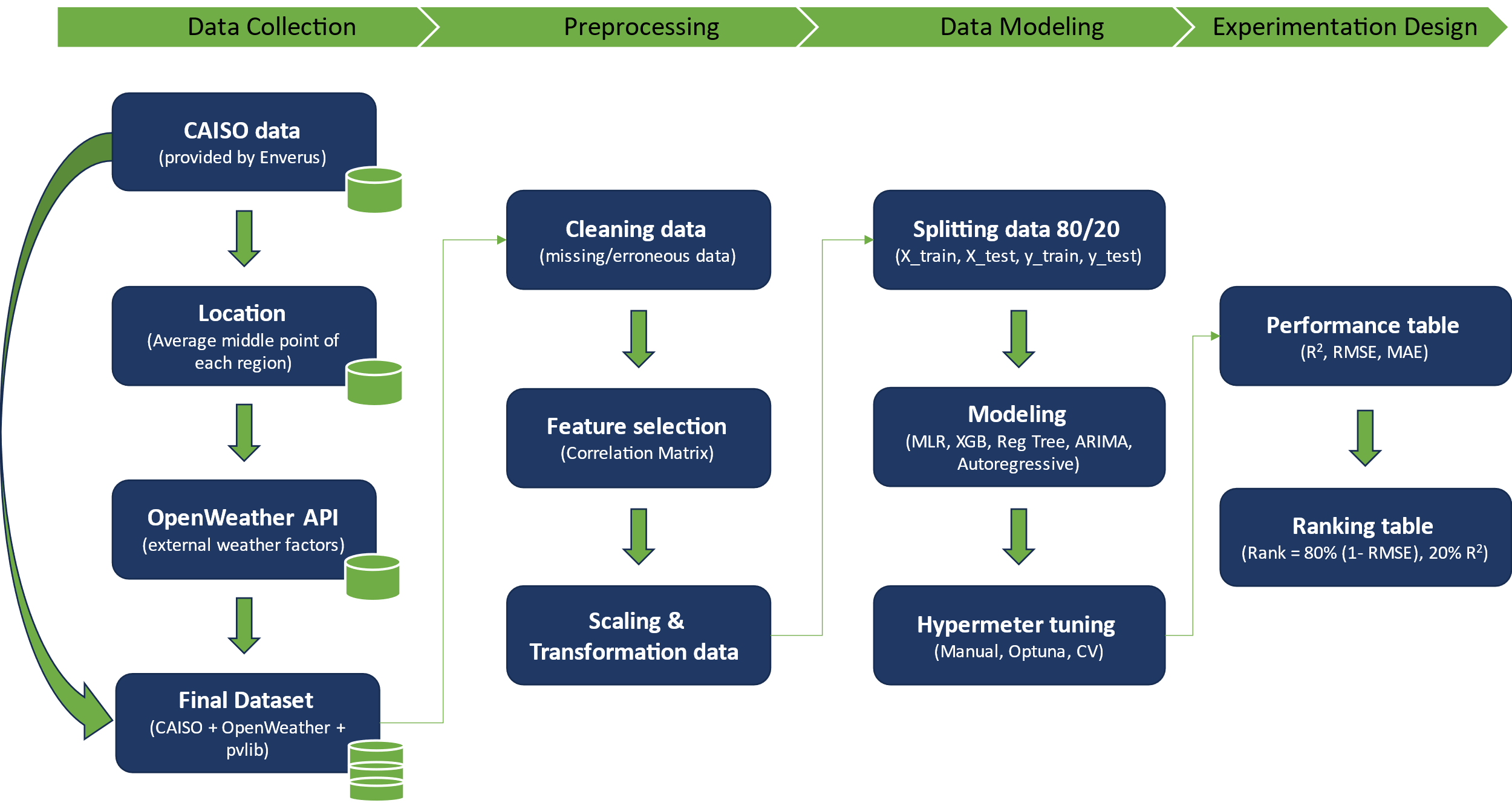
For this study section, we have implemented an extract, transform, and load (ETL) data pipeline, as depicted in Figure 1. The ETL process is crucial for collecting, preparing, and organizing the data to ensure its suitability for analysis and modeling.

Figure 1. ETL Data Pipeline

## 3.1. Data Collection

The data collection process for this project involves gathering relevant information to assess the impact of geographical locations and modeling techniques on the prediction of solar energy generation's performance. We will construct a design matrix containing data combinations from California CAISO and OpenWeather. These data sources will include macro, meso and micro regions in the vicinity of solar location.

Firstly, we will need to understand the supplied data. As part of our collaboration with Enverus, we have been provided a simplified dataset of timestamps and an actual measure of megawatts generated (Appendix A). This 'Actual' measure is based on an average of the three main regions of California. Consequently, as this measure is an average location of the three main regions, we will have to determine a specific location to retrieve future data. In this case, we will take the middle location for each region based on latitude and longitude and will make an average of those coordinates to obtain our central point of California. This process will allow us to coordinate with the data provided by Enverus.

After defining California's middle point location, we will identify and select appropriate data sources that provide environmental factors towards weather factors (temperature, humidity, dust, shading, and wind speed). In this regard, we will use the specific location data from OpenWeather to join it to Enverus' provided data (OpenWeather, 2023)(Appendix A).

Additionally, as part of the requirement from Enverus, we retrieved the measures of Global Horizontal Irradiance (GHI), Direct Normal Irradiance (DNI), and Diffuse Horizontal Irradiance (DHI) from pvlib (Sandia National Laboratories & pvlib python Development Team, 2023). GHI is a measure of the total amount of solar radiation received on a horizontal surface at a given location (HOMER Pro, 2023). It includes both direct solar radiation and diffuse radiation scattered by the atmosphere. The GHI is critical to solar power generation because it provides valuable information about the available solar resources at a given location. By understanding the GHI, solar power system designers can estimate the potential energy output of solar panels and optimize their placement and orientation. GHI data helps determine the feasibility and efficiency of solar projects, allowing for better decision-making when using solar energy to generate electricity.

Indeed, we will attempt to maintain a high level of data quality in this process. To ensure the reliability and integrity of the collected data, we will perform quality assurance with checks. This includes verifying data sources, checking for missing or erroneous data points, and addressing any data inconsistencies or outliers. Indeed, this full dataset represents 26185 rows and 31 columns. Any necessary data cleansing or corrections will be carried out to improve the overall quality of the dataset.

## 3.2. Preprocessing

Data preprocessing is a critical step in preparing the collected data for analysis and modelling for our study.

We will start the explanatory data analysis by cleaning the data. Any missing, erroneous, or inconsistent data points identified during the data collection phase will be addressed. Missing values may be imputed using appropriate techniques, while erroneous or inconsistent data points are corrected or removed based on predefined criteria.

Then, we will proceed with feature selection. We will analyze the full dataset to identify relevant features that contribute significantly to predicting solar energy performance. This step involves establishing a correlation matrix between variables, checking the numerical variables' density, inspecting the categorical variables' distribution, and using domain knowledge to select the most informative features for modeling.

Last but not least, we will proceed with imputation, feature scaling, and transformation. For our missing values, we will proceed with the imputation of the mean value in the numerical variables to avoid the instabilities in the predictions and increase our prediction accuracy (Thomas & Rajabi, 2021). Then, our features will require scaling or transformation to ensure compatibility and optimal performance, as machine learning algorithms can interpret only numerical data (Rebala, Ravi, & Churiwala, 2019). We will proceed with normalizing the numerical data and transforming the categorical data into dummy variable columns with the library Scikit-learn.

## 3.3. Data Modeling

Based on the preprocessed collected data, the data modeling phase focuses on developing models to predict solar energy generation performance. This includes partitioning the dataset into training and test subsets and optimizing model parameters.

Firstly, we will develop and train the models using the preprocessed dataset for each selected modeling technique. Consequently, we will split the preprocessed data into a train and a test set. The test set will represent last four months of data while the train set will be the large remaining part. Then, we will split the dataset with the response variable (y = Actual) and the dependent variables (X). Finally, we will have 4 dataframes corresponding to X\_train, y\_train, X\_test, and y\_test.

Furthermore, based on the study's objectives and the literature review findings, we will select a range of modeling techniques to explore. These models will include traditional statistical models to more advance machine learning models. Indeed, we will use the following models: multi-linear regression (MLR), random forests (RF), eXtreme Gradient Boosting (XGBoost), autoregressive integrated moving average (ARIMA), and Autoregressive model.

In this next step, we will have an experimental set-up phase with the hyperparameters. Finding the ideal collection of hyperparameters for a particular machine learning model is known as hyperparameter optimization (Yang & Shami, 2020). Hyperparameters, such as learning rate, number of layers, or number of units in a neural network, are parameters that are defined before to the training process and are not learnt from the input (Yang & Shami, 2020). Finding the best settings manually can be a time-consuming and tedious process, but it can have a major influence on a model's performance. In order to allow to have some flexibility, we will use different techniques to obtain the best parameters. In this way, our most complete model would be XGBoost with the use of Optuna and cross-validation. Optuna will allow to automatically optimize the hyperparameters by setting a search space dynamically and pruning strategies (Akiba, Sano, Yanase, Ohta, & Koyama, 2019).

## 3.4. Experimentation Design

The experimentation design phase focuses on setting up controlled experiments to compare and contrast different metrics options.

In this section, we will be evaluating model performance using appropriate evaluation metrics with root mean squared error (RMSE) based on the mean square error (MSE), R-squared (R2), and accuracy. To obtain these metrics, we will be using the Scikit-learn library to generate them directly. However, to understand fully the meaning of each of these metrics, we will show their equation:

In regression tasks, the average difference between predicted and actual values is measured using the RMSE (Root Mean Square Error) metric. In order to assess the average error between predicted and observed values in the same unit as the target variable, the square root of the mean of the squared differences is computed (Botchkarev, 2019).

A statistical metric called R-squared (coefficient of determination) shows how much of the variance in the dependent variable can be accounted for by the independent variables in a regression model (Botchkarev, 2019). A better fit of the model to the data is indicated by higher values, which range from 0 to 1.

Our last performance metric for classification models is accuracy. It determines the proportion of all forecasts that were accurate. The result is calculated by dividing the proportion of accurate predictions by the total number of predictions, multiplied by one hundred, and it represents the model's capacity to categorize cases properly (Yin, Wortman Vaughan, & Wallach, 2019).

Lastly, we will design a performance comparison table based on the requirements from Enverus. The performance of the developed models will be compared and contrasted based on the relevant evaluation metrics. Indeed, we will analyze and compare the predictive capabilities of the models based on the equation provided by Enverus. The final ranking of our models will be based on the following equation to determine the best model:

Through this comprehensive methodology, which includes data collection, preprocessing, data modeling, and experimental design, we aim to provide valuable insights into the impact of geographical locations and modeling techniques on the prediction of solar energy generation performance. This approach allows systematic analysis and comparison of different design options, contributing to the advancement of accurate and reliable methodologies in the field of solar energy prediction. In the following section, we will present our findings from the explanatory data analysis.

# Explanatory Data Analysis

For this section of the study, we intend to present the results of the explanatory data analysis of the data collected with the general trends and correlation matrix. These findings helped us to understand and clarify the trends, patterns, and underlying relationships within the data set.

## 4.1. General trends

To understand the data provided by Enverus, we have plotted the "Actual" (average power generation in MW) for the three regions over the entire time period in Figure 2. In this graph, we can clearly see a pattern for each year. The generation of energy tends to be at its maximum in July for the year 2020 to 2022 included. Moreover, this pattern tends to increase each year, which would indicate a real need to focus more on this energy from Enverus for the coming years.

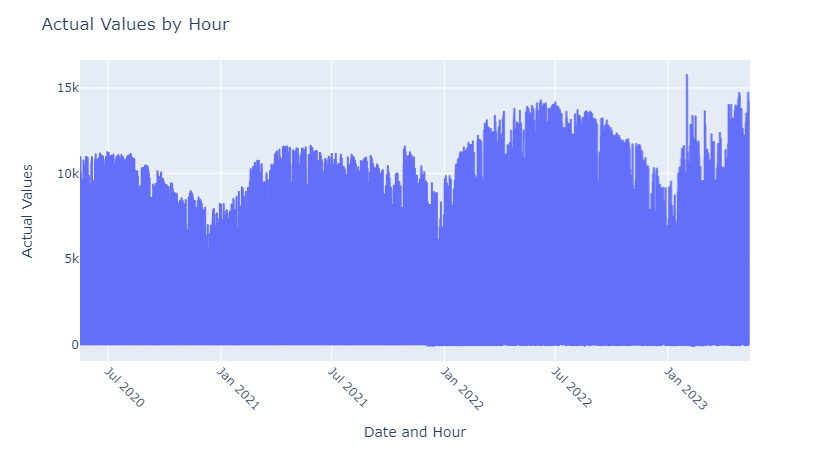


Figure 2. Actual value by hour over the entire period

A graph of a number of colored lines

Description automatically generatedWe then looked at the "Actual" measurement per hour during the day in Figure 3. Here, we confirmed that the generation of energy would generally take place during the daylight hours of the day from 6AM to 8PM.

Figure 3. Actual value by hour through the day

Furthermore, we were wondering to confirm that the generation is generally greater during summer with the long hours of daylight by comparing a week in summer (week 26) and another week in winter (week 51). Based on Figure 4, we can assert that the generation is greater in summer period. In this regard, we can even see a greater difference between summer and winter for the generation of GHI on the right side of Figure 4.

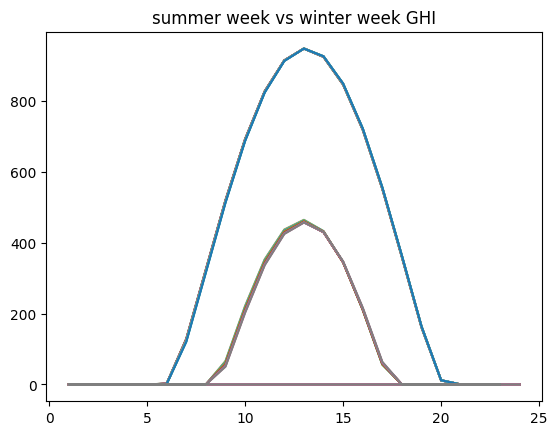
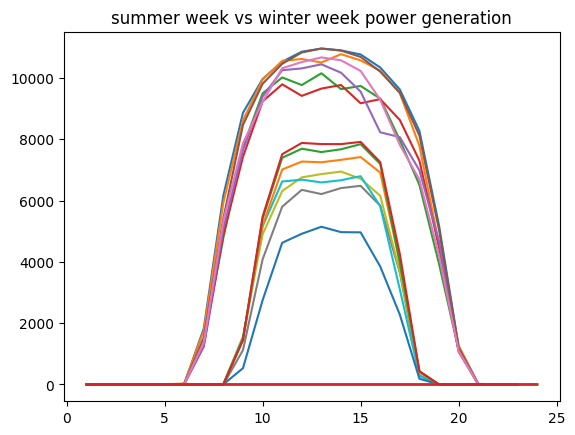


Figure 4. Comparing a week of Actual value in summer and winter

A circular graph with numbers and lines

Description automatically generated Furthermore, we determined the exact location of middle point in California to be of latitude of 37.051548 and longitude of -120.699371. Indeed, the exact address would be CA 152;CA 33, Merced County, California, United States. Determining the exact location helped us unlock further information. Moreover, we illustrated the sun path for our specific location over all the period given in Figure 5. Based on Figure 5, we can also see the hourly solar zenith to be at the hour 13 (1PM). Also, the boundary of the region of the sky that the Sun traverses during a year is marked by the solstices. According to this diagram, there is never a time of the year when the Sun is directly overhead (zenith=0).

Figure 5. Diagram of the analemma

A graph with many colored dots

Description automatically generatedA graph with many colored dots

Description automatically generated Then, we also checked the

Figure 6. Representation of the

A graph of different colored lines

Description automatically generated

Figure 7. Diagram of the analemma

A graph showing a number of data

Description automatically generated

Figure 8. Diagram of the analemma

## 4.2. Feature selection

In this part of the explanatory data analysis, we aimed to filter only the significant columns to be used in our model. In order to do so, we first have a look at the numerical data and then the categorical data.

A green and white graph

Description automatically generated To understand the relationships between the numerical data, we generated its correlation matrix in Figure 9. By calculating the correlation coefficients, we learned more about how strongly one attribute was related to another, such as the different temperature metrics. These strong linear relationships would indicate a high correlation, which could be a sign of redundant or powerful characteristics. Consequently, we will filter out these numerical values.

Figure 9. First Correlation Matrix

After filtering out the high correlation values, we end up with a lower correlation among all numerical values. In Figure 10, we only kept the ‘ghi’ (GHI), which has a high correlation of 92% with our ‘Actual’ values, because it is must have for Enverus and an anticipated correlation for energy generation.

A green and white chart

Description automatically generated

Figure 10. Final Correlation Matrix

To find possible outliers or skewness, we first analyzed the density of the numerical values. Figure 11 shows the distributions of each numeric characteristic, which we calculated as line graphs. This allowed us to assess the central tendency and spread of the data and to identify any potential outliers.

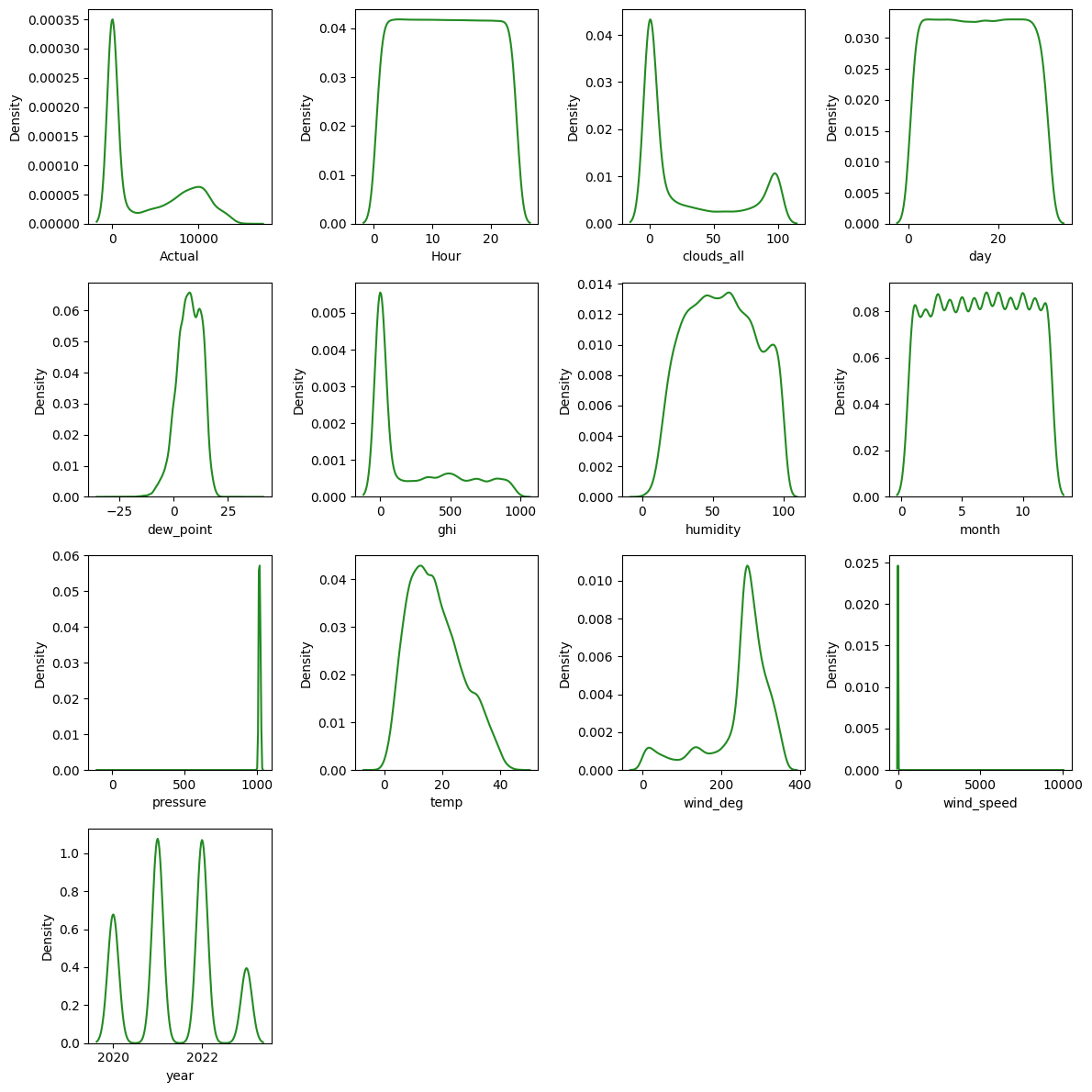


Figure 11. Density of the numerical variables

In fact, the 'Actual' and 'GHI' tend to be slightly skewed to the right. While 'pressure' tends to have a high density in the 1000s, this would suggest fewer data points in that range, indicating that the values in that range are rarer or occur less frequently. Also, "win\_speed" has a high density in the zero mean, which is a distribution of numerical values centered on zero. This indicates that the majority of the data points are clustered around the center and deviate from the mean with relatively little variance.

A graph of different weather conditions

Description automatically generated Regarding the categorical features, we analyzed their distributions with the bar charts in Figure 12. This allowed us to understand the distribution of classes within each categorical feature and identify potential class imbalances or rare categories. Here, we have some highly unbalanced features, like ‘sky is clear’ in ‘weather\_description’ and ‘01n’ & ’01d’ in ‘weather\_icon’, that could introduce bias into the model, while the rare categories might not provide enough information to make accurate predictions.

Figure 12. Distribution of the categorical variables

Regarding our final dataset, we are keeping 12 numerical features and 2 categorical features alongside the response variables. Consequently, our final dataset has 26185 rows and 16 columns before. Indeed, we preferred to stay at this stage of feature selection because of the low number of features and also because certain models contain feature selection at their core. In the following section, we will apply the regression models to our final dataset to generate predictions.

# Results and Visualizations

## 5.1. Linear Regression Modeling

The first model we built was simple linear regression model using only the scaled GHI values. The purpose of this model is to be a reference model for assessing the performance of the upcoming more complex models.. The simple linear regression model was assessed through 5-folds cross validation and achieved overall R2 value of 85.4%. However, these results are optimistically misleading because approximately half of the records in the dataset are within night time with zero power generation. This makes both the training and testing sets imbalanced with roughly half of the target feature values being zero. To overcome this issue, we can eliminate night time hours (zero production hours) which can be easily identified as described during data exploration phase. The night time hours are between 22:00 p.m. to 5:00 a.m. inclusive. After excluding night time hours, the simple linear regression model was reassessed through 5-folds cross validation, and the model achieved overall R2 value of 79.8%. This observation confirms our doubts that keeping zero production hours may result in optimistically misleading results. Then, the dataset was split into training and testing, by allocating the last four months for testing while the earlier part of the data for training. After excluding night time hours, the simple linear regression model was trained using only GHI data, and was able to achieve RMSE of 2379 MW, R2 value of 74.1% on the testing set. A scatter plot of the predicted values versus actual power generation on the testing set suggests that simple linear regression model is not a suitable model for the power generation forecast problem (Figure 13). In addition, the testing error distribution of the testing set does not appears normally distributed which support the earlier observation that simple linear regression model does not fit to the data properly. Figure 14 show a comparison between actual and modeled solar power generation for a 7-day period within the testing set.

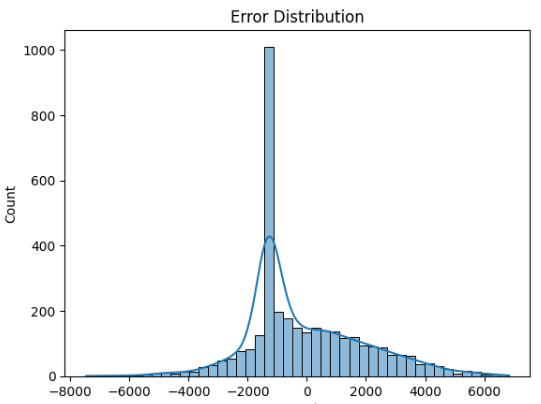
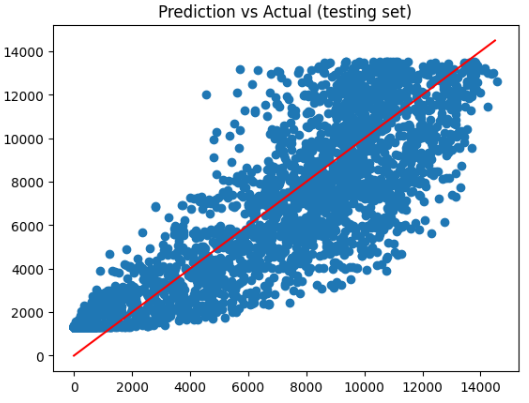


Figure 13. Prediction vs Actual; & error distribution for simple linear regression

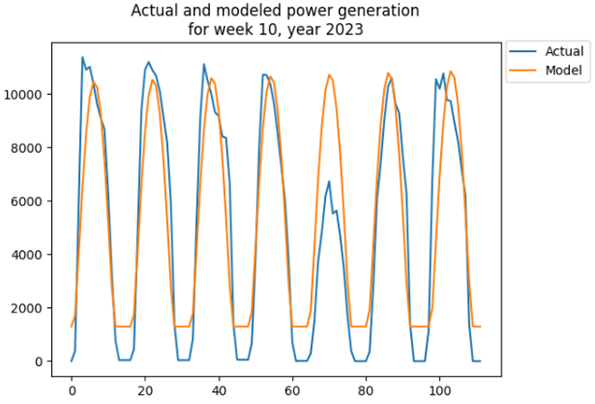


Figure 14. LR - Actual and model power generation in week 10 in 2023

Furthermore, we explored multiple linear regression model using the scaled numeric features of the whole dataset. Categorical features were excluded as their information is already captured in some of the numeric features. Also, we excluded night time hours as described earlier to avoid misleading results. Multiple model fit iterations were implemented while excluding one feature at a time to recognize the significant features. The first model utilizing all numeric features achieved RMSE of 1870 MW, and R2 value of 84.0%, while the final model utilizing only GHI and DNI was able to achieve RMSE of 1897 MW, R2 value of 83.6%. This suggests that both GHI and DNI are significant for forecasting power generation. A scatter plot of the predicted values versus actual power generation on the testing set suggests that MLR model yet is not a suitable model for the power generation forecast problem (Figure 15). In addition, the testing error distribution of the testing set does not appears normally distributed which support the earlier observation that MLR model does not fit to the data properly, although the model’s predictive performance may be fair. Figure 16 show a comparison between actual and modeled solar power generation for a 7-day period within the testing set. Since regression models do not appear suitable, random forest models will be explored next.

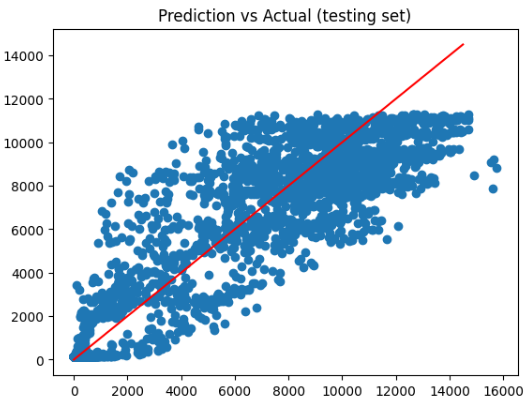
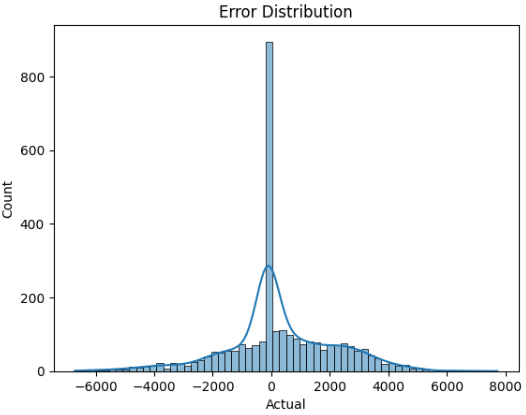


Figure 15. LR scaled - Prediction vs Actual; & testing error distribution for MLR model

A graph of blue and orange lines

Description automatically generated

Figure 16. LR scaled - Actual and model power generation in a winter week in 2021 & 2022

## 5.2. Random Forest

Random forest model was built using the scaled numeric features of the whole dataset. Categorical features were excluded as their information is already captured in some of the numeric features. The initial model was built with 100 trees and using the scaled numeric features. Next, multiple model fit iterations while excluding one feature at a time were implemented and assessed through R2 value and RMSE to recognize the significant features. The assessment revealed that GHI, DNI, cloud cover, humidity, pressure, the week number, and the hour of the day are significant features. Next, the number of trees for the random forest model was optimized based on RMSE and R2 value as shown in Figure 17 below. The optimal number of trees was found to be five, with an overall RMSE of 1666 MW, and R2 value of 87.3% on the testing set. In a similar fashion, the optimal number of samples per leaf was determined to be eleven. Although the model’s predictive performance is relatively good, a scatter plot of the predicted values versus actual power generation on the testing set suggests that random forest may not be a suitable model for the power generation forecast problem (Figure 18). In addition, the testing error distribution of the testing set does not appear normally distributed which supports the earlier observation that random forest model does not have an acceptable fit to the data. Figure 19 shows a comparison between actual and modeled solar power generation for a 7-day period within the testing set. The model appears to be consistently underestimating peak hours energy output. For this reason, we explored a different approach by creating two random forest models: one for peak hours, and another for off-peak hours. However, this approach resulted in RMSE increase to 2492 and R2 value decrease to 71.6%, suggesting that aggregating model approach may not be the optimal way to go. For this reason, time series modeling will be explored afterward.

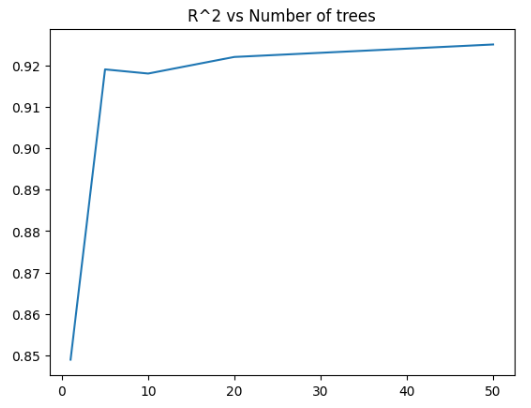
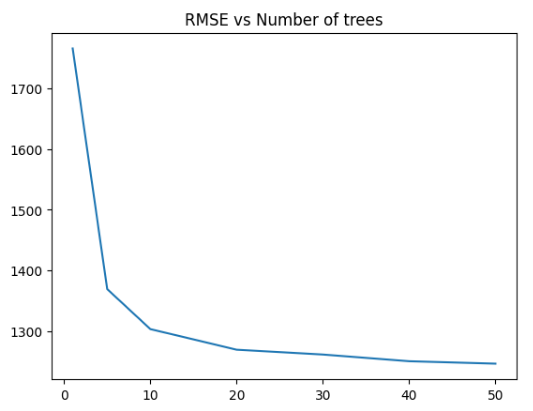


Figure 17. RF - RMSE & R2 vs Number of Trees

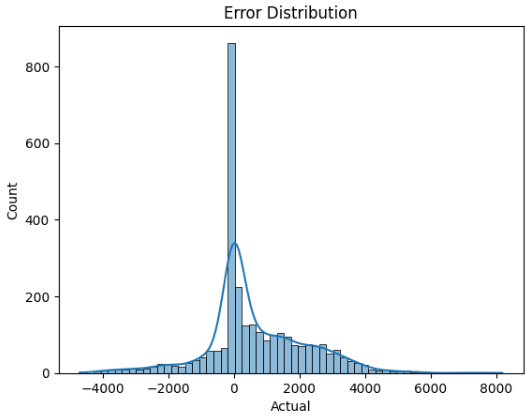
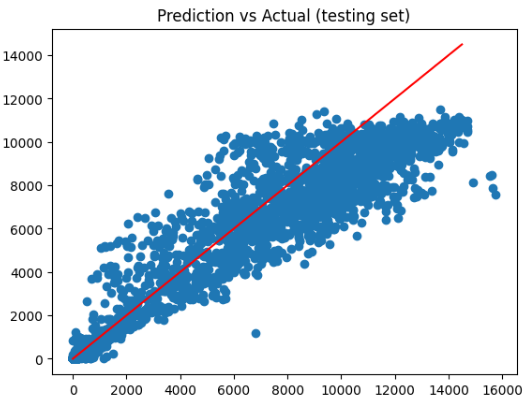


Figure 18. RF - Prediction vs Actual; & error distribution for random forest model

A graph of blue and orange lines

Description automatically generated

Figure 19. RF - Actual and model power generation in a summer week in 2021 & 2022

## 5.3. Time Series Modeling

Time series models were explored on the actual solar power generation without excluding night time. The auto\_arima function was utilized to optimize the parameters p, d, q, comprising the order of the model. The function auto\_arima suggested that the optimal order of Arima is (1,1,2). However, using this model order to forecast one week ahead revealed that it was poor as shown in Figure 20. Upon exploration, we noticed that only the autoregressive part of ARIMA model was the significant parameter, meaning that ARIMA model simplifies to autoregressive model. To identify the optimal number of lags in autoregressive model, we plotted the partial autocorrelation in the whole dataset as shown in Figure 20, and determined the optimal number of lags to be 50 hours. The autoregressive model was trained on 720 hours, to estimate the parameters associated with each of the 50 lags, in order to forecast the future 72 hours (3 days). The autoregressive model achieved an overall RMSE of 1900 MW, and R2 value of 76.1% which is relatively low. However, upon inspection, we noticed that the model forecast tends to lag ahead of the actual output as shown in Figure 21. For this reason, we decided to manually shift the forecast one hour earlier. This simple trick resulted in outstanding improvement of the autoregressive model, leading RMSE to decline down to 1055 MW and R2 increase up to 89.5%. Figure 22 shows the same forecast period used in Figure-11, after applying the shifting trick. In addition, the testing error distribution for the model appears approximately normally distributed as shown in Figure 23. Consequently, autoregressive models may be suitable for the power generation forecast problem, even though they lack explainability.

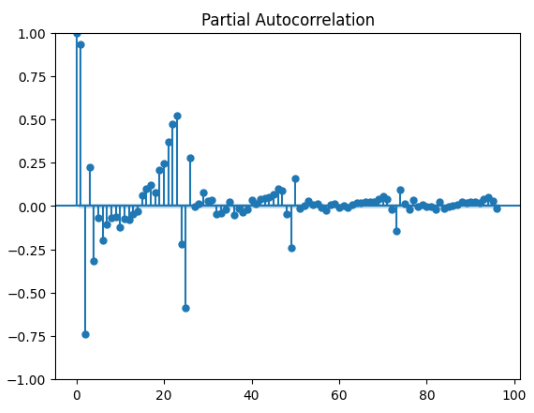
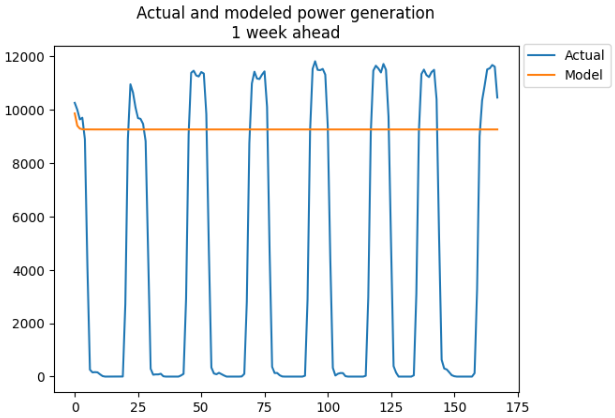


Figure 20. ARIMA - Actual and model power and Autoregressive partial autocorrelation

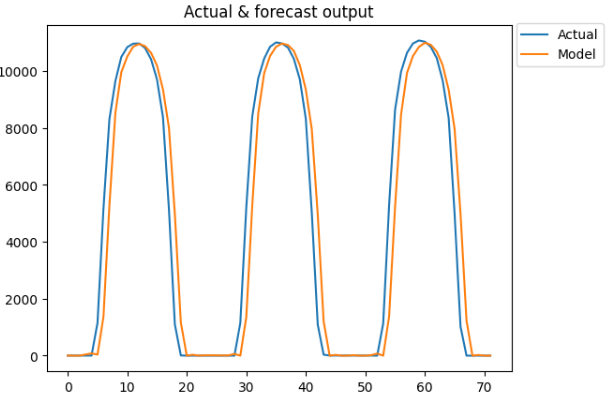
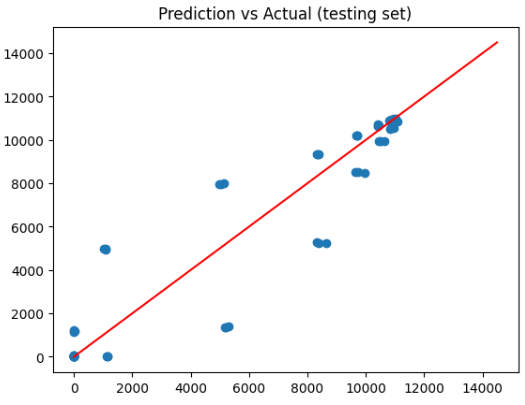
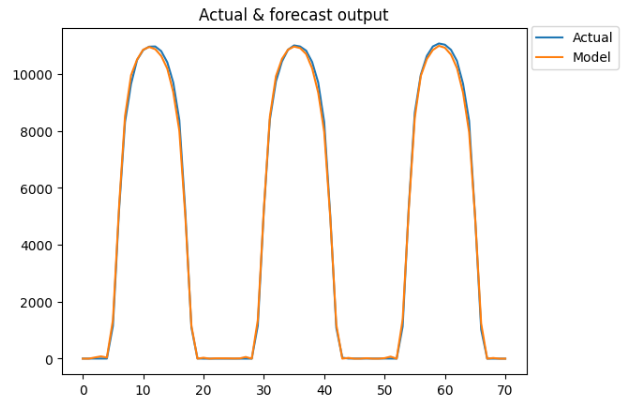
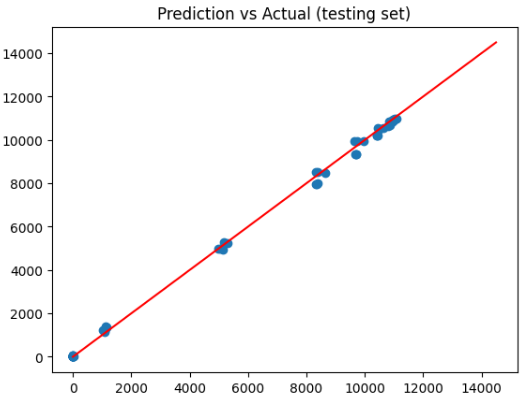


Figure 21. Autoregressive - Autoregressive model forecast over a 3-day period



A graph of error distribution

Description automatically generatedFigure 22. Autoregressive - Autoregressive model forecast over a 3-day period after shifting trick.

Figure 23. Autoregressive - Testing error distribution for shifted autoregressive model

## 5.4. XGBoost

For our last model, we are using XGBoost because it is well known for its ability to handle complicated data sets and deliver powerful results. Its widespread use is due to the sophisticated ensemble learning approach, which combines the predictions of multiple separate models, or weak learners, to produce a robust final model (Srinivas & Katarya, 2022). XGBoost effectively handles a variety of feature types thanks to parallel processing and tree-based learning techniques, making it suitable for a wide range of fields. Also, the regularization algorithms used by XGBoost ensure reliable predictions by preventing overfitting and improving generalization (Srinivas & Katarya, 2022).

Firstly, we ran the model without any optimization on the split final dataset. In Figure 26, we have the feature importance displayed. As expected, ‘ghi’ and ‘day’ are the top ones due to their high correlation. Then, the rest of the features hold some importance, but it may change after optimization of the hyperparameters.

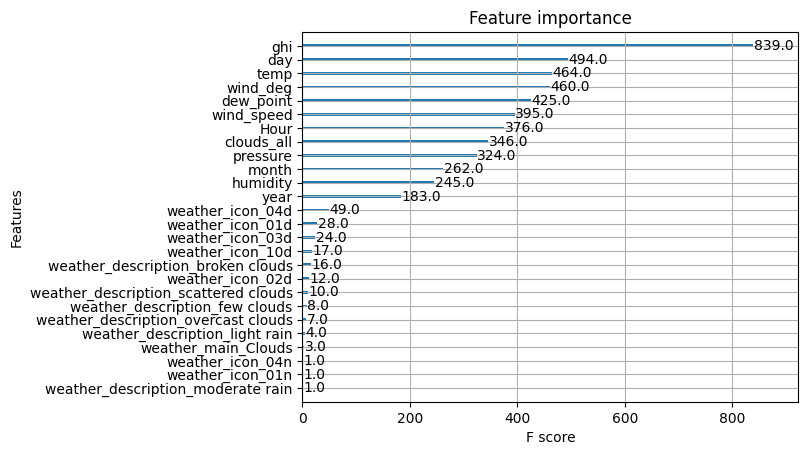


Figure 26. XGB - Feature importance on non-optimized model

However, no model promises perfect results under all circumstances. Proper evaluation, tuning, and handling of potential problems such as overfitting are critical. Model performance is influenced by the handling of class imbalances and hyperparameter tuning. To effectively search the hyperparameter space and determine the ideal configuration for XGBoost, Optuna uses state-of-the-art optimization methods, including Bayesian optimization and tree-structured Parzen estimators (Akiba, Sano, Yanase, Ohta, & Koyama, 2019; Srinivas & Katarya, 2022). It effectively searches for hyperparameters that lead to improved performance by judiciously balancing exploration and exploitation.

We can automate the hyperparameter tuning process for XGBoost using Optuna's capabilities, which will save a lot of time and effort. Optuna intelligently samples hyperparameter combinations and evaluates them using specified objective functions or evaluation metrics, as opposed to manually tweaking hyperparameters and evaluating model performance. Because Optuna provides a simple API to specify the search range for hyperparameters and an interface to communicate with the XGBoost training process, its integration with XGBoost is smooth (Akiba, Sano, Yanase, Ohta, & Koyama, 2019). The fact that it supports both discrete and continuous hyperparameters makes it possible to perform a thorough search over a wide range of values.

In addition, Optuna provides features such as early pausing, trimming, and parallel execution that further speed up and increase the effectiveness of the hyperparameter optimization process. It intelligently terminates unproductive experiments early and efficiently distributes computing resources. As shown in the Figure 27, Optuna quickly found the best objective value for our model.

A graph with blue and red dots

Description automatically generated

Figure 27. XGB Optuna - Optimization History Plot

Optuna provides a Parallel Coordinate plot which helps decipher and evaluate the results of hyperparameter tuning. It provides a clear and simple perspective that makes it easy to identify high performing hyperparameter setups by displaying each trial as a line and mapping the hyperparameters on multiple axes for our model. Indeed, Figure 28 clearly shows a repetitive pattern that would work the best for our dataset.

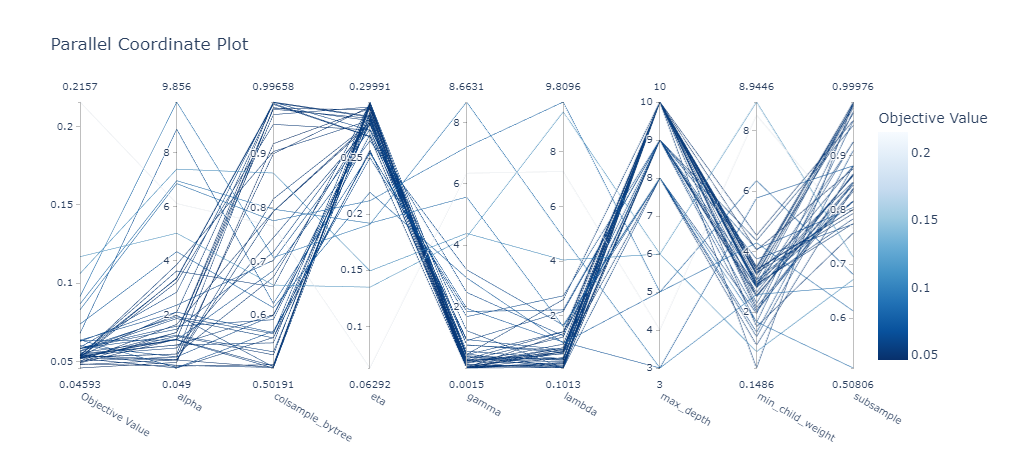


Figure 28. XGB Optuna - Parallel Coordinate Plot

As shown in Figure 29, the final hyperparameters would highly rely on the eta parameter with a significance of 67%. The shrinkage applied to each tree's contribution in the ensemble is controlled by the eta parameter, which impacts the model's convergence rate and overall performance (Akiba, Sano, Yanase, Ohta, & Koyama, 2019).

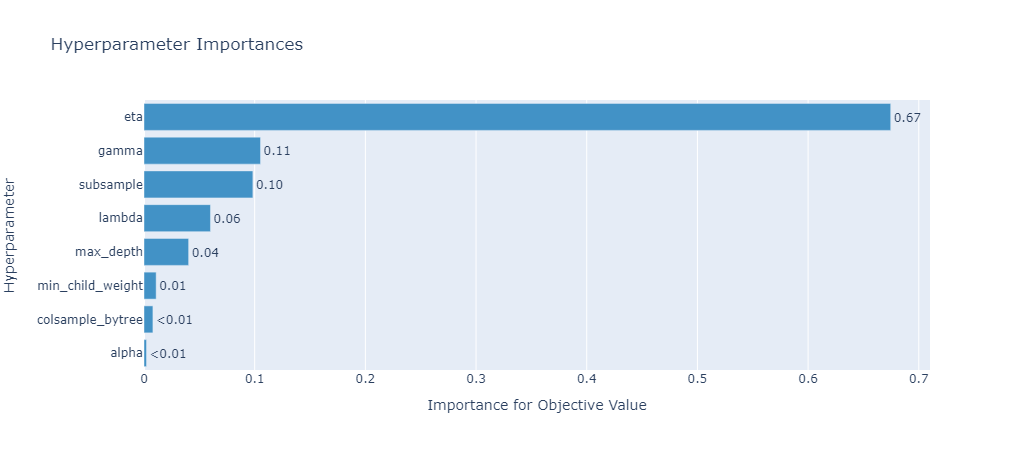


Figure 29. XGB Optuna - Hyperparameter Importance

After using Optuna for our hyperparameter tuning, we can observe a shift in the feature importance (Figure 30). ‘ghi’ and ‘day’ remain high, but ‘month’ finally increased, which would be logical as this is also a timestamp and indicator of the season too for energy generation. Indeed, Figure 30 illustrates a better importance of the features when modeling for the prediction of energy generation.

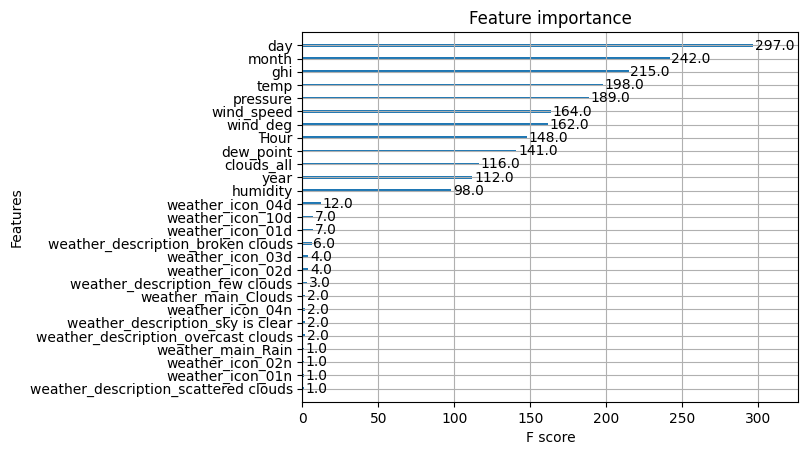


Figure 30. XGB Optimized - Feature importance on optimized model

Then, we have plotted our predictions against the actuals in Figure 31. We can observe that our model is pretty conservative in comparison to what actually happened.

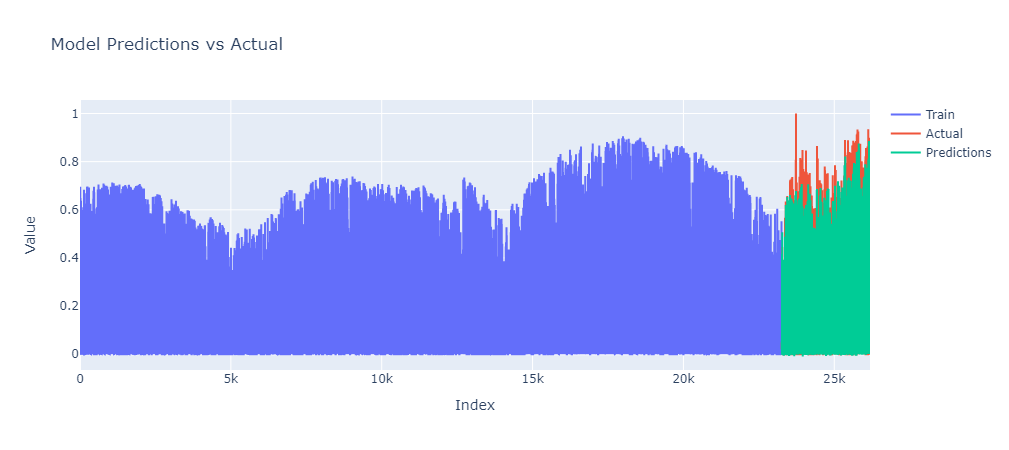


Figure 31. XGB Optimized - Model Predictions vs Actual

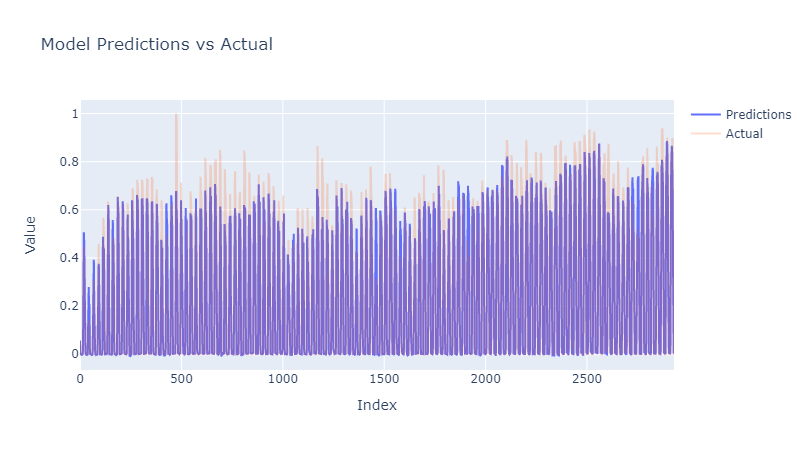
Now, we will have a look at the predictions only to understand what happened. Figure 32 represents the last 4 months of data, where we tested our predictions. Once we slightly gray out the actual, we can better see the predictions. Indeed, we can see that our model is quite conservative, but it is following most of the trends from the actual.

Figure 32. XGB Optimized - Model Predictions vs Actual focus on predictions period

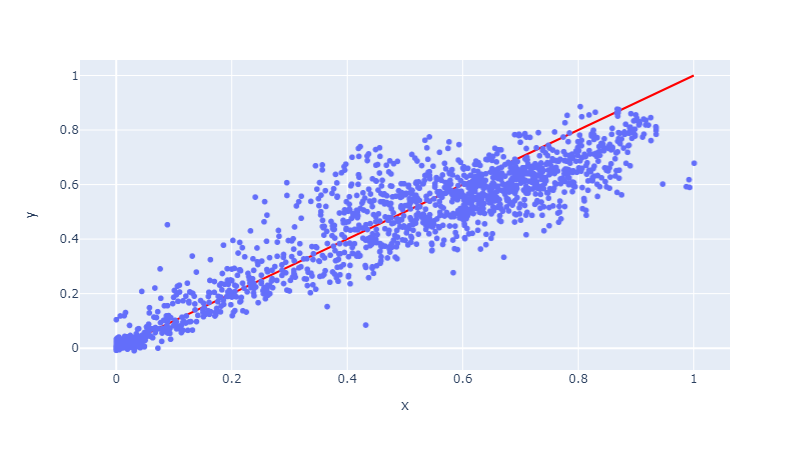
 Based on Figure 33, we observed that our predictions against the actual still clearly form a line trends which would indicate the validity of this model.

Figure 33. XGB Optimized - Model Predictions vs Actual

# Conclusion and discussion

This page will be almost 1 page and half long.

# References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization. *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2623–2631. doi:https://doi.org/10.1145/3292500.3330701

Aksoy, N., & Genc, I. (2023). Predictive models development using gradient boosting based methods for solar power plants. *Journal of Computational Science, 67*(101958), 1-10. doi:https://doi.org/10.1016/j.jocs.2023.101958

Botchkarev, A. (2019). Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology. *Interdisciplinary Journal of Information, Knowledge, and Management*, 045-076. doi:https://doi.org/10.48550/arXiv.1809.03006

California ISO. (2023). *California ISO Open Access Same-time Information System.* Retrieved from California ISO: http://oasis.caiso.com/mrioasis/logon.do?reason=application.baseAction.noSession

Chikate, B. V., & Sadawarte, Y. (2015). The factors affecting the performance of solar cell. *International journal of computer applications, 1*(1), 1-5. doi:0975 – 8887

Enverus. (2023). *Intelligent Connections*. Retrieved from Enverus: https://www.enverus.com

Hobbs, B. F., Zhang, J., Hamann, H. F., Siebenschuh, C., Zhang, R., Li, B., . . . Zhang, S. (2022). ISO, Using probabilistic solar power forecasts to inform flexible ramp product procurement for the California. *Solar Energy Advances, 2*(100024), 1-11. doi:https://doi.org/10.1016/j.seja.2022.100024

HOMER Pro. (2023). *Global Horizontal Irradiance (GHI)*. Retrieved from HOMER Pro: https://www.homerenergy.com/products/pro/docs/3.11/global\_horizontal\_irradiance\_ghi.html

Miao, S., Ning, G., Gu, Y., Yan, J., & Ma, B. (2018). Markov Chain model for solar farm generation and its application to generation performance evaluation. *Journal of Cleaner Production, 186*(1), 905-917. doi:https://doi.org/10.1016/j.jclepro.2018.03.173

OpenWeather. (2023). *Weather API.* Retrieved from OpenWeather: https://openweathermap.org/api

Rebala, G., Ravi, A., & Churiwala, S. (2019). *Machine Learning Definition and Basics. In: An Introduction to Machine Learning.* Worldwide: Springer, Cham. doi:https://doi.org/10.1007/978-3-030-15729-6\_1

Sandia National Laboratories & pvlib python Development Team. (2023). *User Guide*. Retrieved from pvlib: https://pvlib-python.readthedocs.io/en/stable/user\_guide/index.html

Singh, A. K., & Singh, R. R. (2021). An overview of factors influencing solar power efficiency and strategies for enhancing. *Innovations in Power and Advanced Computing Technologies (i-PACT), 1*(1), 1-6. doi:10.1109/i-PACT52855.2021.9696845

Thomas, T., & Rajabi, E. (2021). A systematic review of machine leaning-based missing value imputation techniques. *Data Technologies and Applications, 55*(4), 558-585. doi:https://doi.org/10.1108/DTA-12-2020-0298

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing, 415*(1), 295-316. doi:https://doi.org/10.1016/j.neucom.2020.07.061

Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the Effect of Accuracy on Trust in Machine Learning Models. *Association for Computing Machinery*, 1–12. doi:https://doi.org/10.1145/3290605.3300509

# Appendix A

Data provided:

OpenWeather data: